1 Introduction

In this project we are going to explore how changes in the forest carbon stock correlates with the surface temperature, do we even have a correlation in both or not.

2 Research Question

Is there a correlation between changes in forest carbon stocks and surface temperature changes?

3 Data Sources

3.1 Forest and Carbon Dataset

- **Metadata URL:** IMF Forest Data This dataset provide forest areas and carbon stock in forest and land area from 1992 to 2020, this is open source data from International Monetary Fund (IMF)

3.2 Annual Surface Temperature Change Dataset

- **Metadata URL:** IMF Surface Temperature Data This data Provides Mean surface temperature change from 1961-2021 for different countries, data is open source provided by International Monetary Fund (IMF).

3.3 Data Structure and Quality

- Both Datasets are in CSV format and after transformation stored in a SQLite Database, the dataset is also majorly complete and consistent and accurate.
- **Tables:**
  - *Annual_Surface_Temperature_Change:* Columns for *Country*, *Indicator*, and annual data from 1961 to 2022.
  - *Forest_and_Carbon:* Columns for *Country*, *Indicator*, and annual data from 1992 to 2020.
  - *Temp_Change_Diff:* Annual differences in temperature change for each country.
  - *Carbon_Stocks_Diff:* Annual differences in carbon stocks for each country.

3.4 Licenses

datasets can be used and distributed with proper citation. IMF Data Terms of Use. This open data policy facilitates academic and public research.

4 Data Pipeline

4.1 Description

The pipeline fetches the data from the URLs first, then it cleans the data by removing those that that contains missing values and convert the required column to numeric format,then it transforms the data for analysis by calculating the annual differences in temperature and carbon stocks and finally in last stores the data in a SQLite database using Python, pandas, and SQLite. it basically follow ETL(Extract, Transform, Load) Pipeline standard.

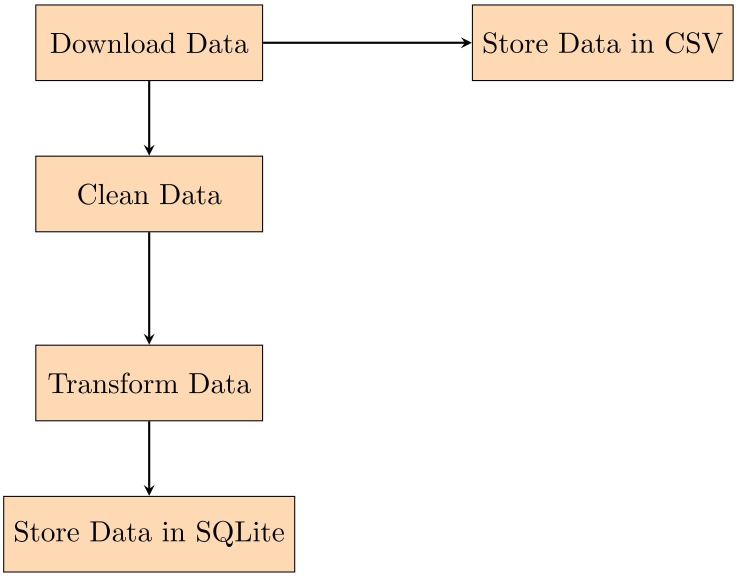

Figure 1: Data Pipeline Diagram

4.2 Transformation and Cleaning Steps

we first downloaded the data and loaded it into the Pandas dataframe then we removed the columns that are not so important for data simplification and then we removed the records that had some missing values, converted the required column types to numeric for a better analysis result, made the country colum an index and finally Computed Yearly changes to find out the trends over time.

4.3 Problems Encountered

- So while working with the data i faced majorly 2 common problem, first removing the rows that contains NaNs to maintain the quality of the data and secondly the columns that i need to convert into numeric i had to make sure that those are correctly formated, so these are some problems that i encountered while working with the data.

5 Result and Limitations

5.1 Output Data

The pipeline output data has:

- Annual temperature change data for several countries.
- Forest and carbon stock data for several years.
- Annual differences in temperature changes and carbon stocks, for detailed trend analysis.

5.2 Data Format

Reason for Choosing SQLite:

- **Efficiency:** Efficient storage and querying.
- **Portability:** Easily transferable and shareable.
- **Ease of Use:** Simple to use with with pandas for data manipulation.

5.3 Critical Reflection and Issues

- **Strengths:** the data is Clean and consistent and can be queried perfectly.
- **Potential Issues:**
  - **completion of data:** after Dropping some rows that had some missing values can or may lead to loss of some important info.
  - **Linear Assumptions:** Annual differences assume linear changes, which sometimes overlook complex patterns.
  - **Geographical and Temporal Coverage:** data is limited to certain regions or periods which may affect trend analysis.
  - **Analysis Limitations:** The linear assumption in annual differences might oversimplify real world dynamics.
- **Future Adaptations:** The pipeline might need modifications to coupe up with changes in data sources or to integrate additional datasets for a more comprehensive analysis.